DOCUMENT RESUME

ED 148 842 TM 006 727

AUTHOR Knapp, Joan; Sharon, Amiel
TITLE A Compendium of Assessment Techniques.
INSTITUTION Cooperative Assessment of Experiential Learning, Columbia, Md.
SPONS AGENCY Carnegie Corp. of New York, N.Y.
PUB DATE Jan 75
NOTE 51p.
AVAILABLE FROM Cooperative Assessment of Experiential Learning, American City Building, Suite 403, Columbia, Maryland 21044 ($3.50)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *College Students; Essay Tests; *Evaluation Methods; Higher Education; Informal Assessment; Interviews; Job Skills; *Learning Experience; Measurement Techniques; Objective Tests; *Performance Tests; Self Evaluation; Simulation; Situational Tests; *Student Evaluation; *Work Experience; Work Sample Tests
IDENTIFIERS *Experiential Learning; Portfolios

ABSTRACT

A wide variety of techniques are appropriate for evaluating different types of experiential learning. This monograph includes sections on performance tests, simulations, interviews, ratings, self evaluation, and product assessment. Each section orients the reader to one of these general types of assessment and then provides brief illustrations of different applications. A number of references concerning each assessment method are cited. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

CAEL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

cael

ED148842

TM006 727

A COMPENDIUM OF ASSESSMENT TECHNIQUES

JOAN KNAPP
AMIEL SHARON

COOPERATIVE ASSESSMENT OF EXPERIENTIAL LEARNING

COOPERATIVE ASSESSMENT OF EXPERIENTIAL LEARNING (CAEL)

is a joint project of Educational Testing Service and a group of colleges and universities with initial funding by the Carnegie Corporation:

Antioch College
California State University and Colleges
(San Francisco State University)
Community College of Vermont
El Paso Community College
Empire State College
Florida International University
Massachusetts State College System
(Framingham State College)
Minnesota Metropolitan State College
New College (University of Alabama)
Thomas A. Edison College

Please address inquiries to: CAEL
Educational Testing Service
Princeton, NJ 08540

A COMPENDIUM OF ASSESSMENT TECHNIQUES

This compendium was prepared by Joan Knapp and Amiel Sharon, Educational Testing Service. A variety of source material, including unpublished reports, journal articles, books, and documents in the ERIC system, was used in this paper. The authors extend a special note of acknowledgement to three unpublished reports that provided information and suggestions regarding the application of various non-traditional assessment techniques:

Assessment Handbook (in draft), by Aubrey Forrest;

Guidelines for Developing and Evaluating Outcomes of a Comprehensive Training Program, by Lois A. Crooks;

The Selection and Development of Performance Measures for Assessment Center Programs, by Lois A. Crooks.

Thanks are also due to the CAEL Implementation Committee for their review of the manuscript and for numerous useful suggestions.

Cooperative Assessment of Experiential Learning
January 1975

Copyright © 1975 • Educational Testing Service • All Rights Reserved.

CONTENTS

I. Introduction 5

II. Performance Tests 7

III. Simulations 13

IV. Assessment Centers 25

V. Essay Examinations 29

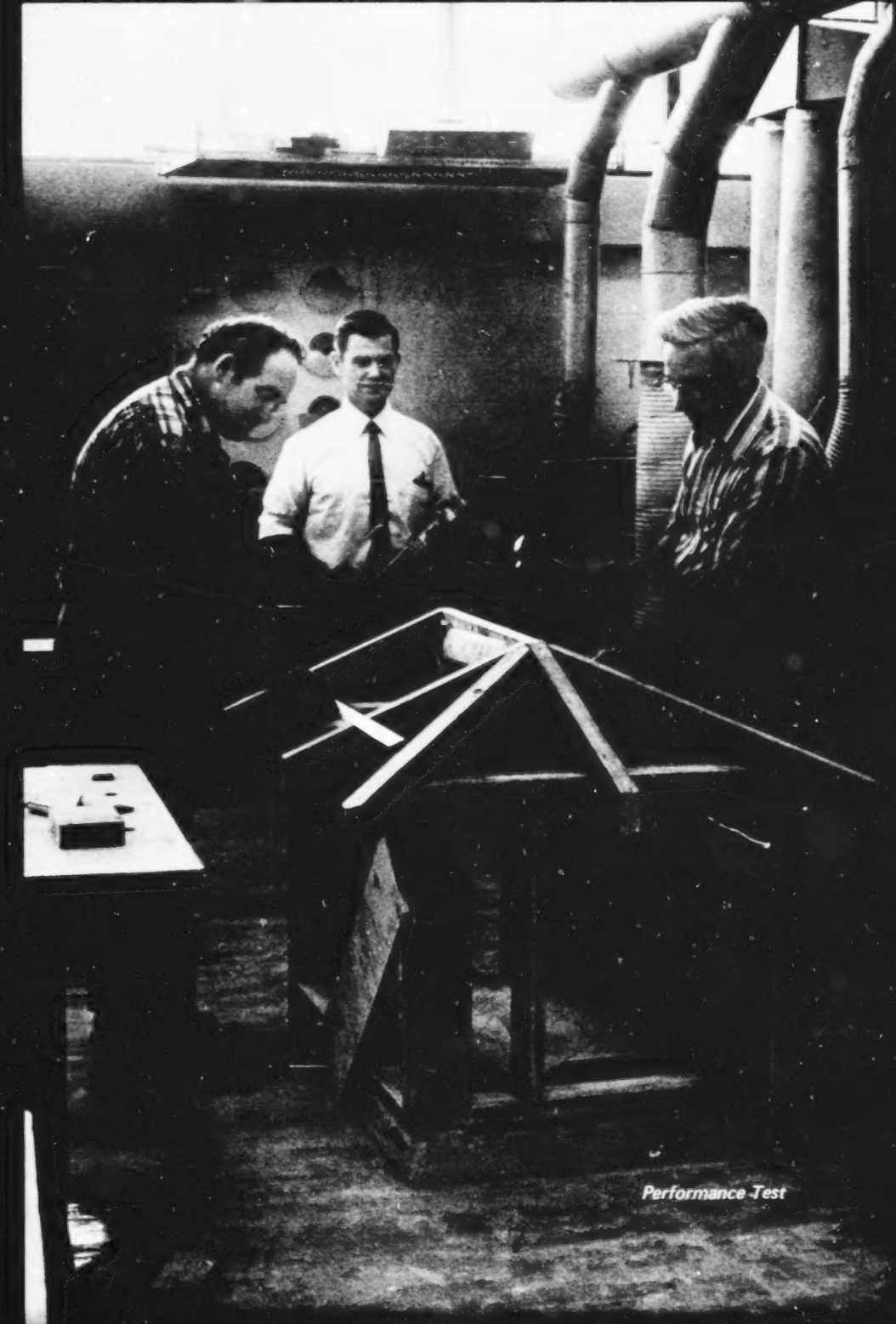
VI. Objective Written Examinations 33

VII. Interviews 37

VIII. Self-Assessment 43

IX. Ratings 47

X. Product Assessment 49



Performance Test

I. INTRODUCTION

Because experiential learning, in the sense employed in the CAEL Project, often differs from traditional classroom learning, its assessment and evaluation require the consideration of measurement techniques that have not been widely employed in higher education. The aim of this compendium is to provide a brief introduction to techniques that can be used to assess student learning "by doing"—the kind of learning that takes place when an individual participates in a variety of social, artistic, political, work, or cross-cultural activities off the college campus.

This is not to say, of course, that the assessment techniques discussed in this paper are irrelevant to traditional forms of learning on college campuses. Indeed, higher education is much in need of improved methods of assessing student achievement—valid and equitable methods that take fully into account the individual character of students and their learning experiences. Development of improved means of assessing experiential learning should have corresponding benefit for higher education generally.

Several of the measurement procedures to be described have been traditionally used in business and industry rather than in an educational context. Although the objectives of assessment in business and in higher education are likely to be different, often the same techniques can be used to evaluate the competence of an individual in either situation.

This publication will provide a relatively brief discussion of a number of assessment techniques, with sufficient detail to allow the reader to determine whether a particular technique might be applicable to his own situation. Although not a procedural manual showing how to develop and administer assessment techniques, it does provide sufficient information and references for the reader who wants to know more about a particular technique.

The major categories of assessment techniques described are ordered on the basis of two factors. The first is the degree of realism in the assessment situation. The first five major categories range from those techniques that have a strong element of realism (for example, performance tests) to those that present a highly artificial and restricted assessment situation (such as written objective examinations). Between these extremes are the simulation techniques, assessment centers, and essay examinations, all of which are moderately realistic. The other factor that distinguishes one category from another is the source of the data being evaluated. The student is the source of information in the interview and self-assessment techniques. However, in ratings and product assessment, someone other than the student provides the basic data for evaluation.

The discussion of each major assessment technique contains the following sections:

Description identifies the mode of assessment (for example, the kinds of stimuli presented, the kinds of responses given), the kinds of materials used, how the technique is administered, the kinds of competencies assessed, and the major objective and setting of the assessment.

Illustration presents a specific example of how the technique could be used to assess the learning of a candidate with particular experiences.

Development outlines the procedure for developing the technique and discusses any special problems involved.

Technical considerations describes the technique's psychometric properties, including its objectivity, reliability, and validity.

Evaluation estimates the appropriateness, feasibility, and cost of the technique in assessing experiential learnings and suggests some kinds of learning for which the technique is appropriate.

References presents books, journal articles, published tests, and other sources that describe the use of the technique or evaluate it.

II. PERFORMANCE TESTS

A strong element of realism exists in performance tests because they resemble the situation in which a specific kind of experiential learning is applied. Frequently, a performance test is nothing but a work sample requiring the accomplishment of specific tasks in a controlled setting. Because it is often impractical or dangerous to reproduce the real job situation or to test with actual equipment, the critical operations and tasks of the job are reproduced by special equipment or procedures called simulators. Perceptual and motor tasks such as typing or operating radar equipment as well as conceptual tasks such as troubleshooting or interacting with other persons can be simulated. Since written and oral communication skills are essential to many job activities, some performance tests are primarily verbal in nature. Others emphasize manipulative skills.

A student's performance can be observed in the laboratory under controlled conditions or in the environment in which it occurs. Unobtrusive observation of performance in a natural setting is desirable when the subject's behavior is likely to be altered significantly by his awareness of being observed. Unobtrusive observation can be facilitated by hidden hardware such as a microphone or television camera or by the analysis of available institutional records.

Performance tests are less efficient and more costly to administer than written examinations. They often require one examiner per examinee, special testing equipment, observing and recording hardware, and a relatively long testing time.

The characteristics of two major types of performance tests—the work sample and unobtrusive observation—will be described in the following paragraphs. Performance tests with somewhat less realism than work samples and actual job performance but more than conventional written examinations are described in the section on simulations, beginning on page 13.

Work Sample

A work sample is a test that attempts to reproduce all or an important part of the actual operations and tasks of a job. The test utilizes the actual equipment used on the job, or, if such equipment is hazardous, simulators. Sometimes, the test involves certain difficulties typical of the job which the examinee must overcome. Thus, an electronics technician may be required to check out circuits and to identify and repair malfunctions on a piece of simulated equipment. Some of the realism of the work setting may be sacrificed, but the critical job elements are present and the test is readily recognized as a realistic representation of the tasks one would encounter on the job. Familiar work sample tests that employ the actual equipment of the job include the road test required of driver license applicants, typing proficiency tests, and foreign language proficiency tests.

Illustration. A veteran of the Air Force trained as a communication equipment specialist wants to receive credit towards an associate degree in electronics

technology. To demonstrate his proficiency in the manipulative skills and judgments essential to the occupation, he takes a work sample test involving various standard recording and communications equipment. The examinee diagnoses problems and makes repairs using various tools such as pliers and screwdrivers. The performance is evaluated by an expert examiner on the basis of speed of diagnosis, efficiency of repair, and adequacy of the final product.

Development. Work samples should be developed on the basis of performance objectives derived from a careful job analysis. The tasks selected for the work sample must be representative of the job and sufficient time allowed so that the scores will be reliable. The appropriate testing conditions, including tools, equipment, and materials required for the specified tasks, must be provided. Detailed instructions on what is to be done and the basis for evaluating performance must be given to the examinee. Special instructions and forms must be developed for the test administrator so that he can record and evaluate the performance. The relative weight to be given to the process and the product of performance will depend on the objectives of the test and the nature of the task. Preliminary tryouts of the test should be made on experienced individuals, to detect problems or difficulties. Although many work sample tests are criterion-referenced, norms on experienced and inexperienced workers could be developed as an aid to score interpretation.

Technical considerations. A work sample is useful in direct proportion to how well it represents the actual work on the job. The test cannot assess all skills because some skills (such as managing a business) are not reducible to a test situation without extensive simulation. Several sources of error or unreliability tend to reduce the validity of the technique. These include the lack of uniform testing conditions, using equipment with which some candidates may be unfamiliar, and using inexperienced or untrained observers.

Evaluation. Work samples are used extensively for job selection and training where high fidelity between the work sample and job is possible. The tests are therefore appropriate for the assessment of work experiences, especially for students majoring in vocational-technical subjects. They are also appropriate for the assessment of proficiency in a foreign language, art, music, sports, and vehicle operation. The high cost of developing and administering the tests may deter their use in higher education. Nevertheless, work samples are an essential technique in the assessment of some skills acquired through experiential learning.

References

Boyd, J.L., Jr., and Shimberg, B., *Handbook of Performance Testing, A Practical Guide for Test Makers.* Princeton, NJ: Educational Testing Service, 1970.

This book describes how to plan, develop, and grade work sample tests, and the types of documents needed for their administration. It provides a portfolio of tests, such as those for machinist, cosmetologist, dental hygienist, and bench woodworker.

Fitzpatrick, R., and Morrison, E.J., "Performance and Product Evaluation," in Thorndike, R.L., ed., *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education, 1971, 237-270.

This chapter describes work sample tests and distinguishes them from other kinds of simulated testing situations.

National Occupational Competency Testing Institute (NOCTI) Examinations

A series of written and performance examinations in 24 major occupations administered nationwide on a regular basis. The major purpose of the examinations is to grant academic credit to individuals with work experience who desire to become vocational teachers. For information write to: Program Director, NOCTI, Educational Testing Service, Princeton, NJ 08540.

Panitz, A., and Olivo, T.C., *Handbook for Developing and Administering Occupational Competency Testing.* New Brunswick, NJ: Rutgers University, 1971.

This handbook provides a step-by-step description of the development, administration, scoring, and analysis of occupational work sample tests.

Sharon, A.T., *Planning the Development of Measurement and Evaluation Services for Use in Occupational Programs at Postsecondary Institutions.* (PR-74-16) Princeton, NJ: Educational Testing Service, 1974.

This publication describes the needs of postsecondary institutions for occupational competency measures. The focus is on translating work experience into degree credit in the fields of accounting, agribusiness, day care, electronics technology, data processing, management, police science, and secretarial science.

Unobtrusive Observation

Unobtrusive assessment of performance occurs when the examinee is not aware of being observed. The observer may be hidden in a crowd, may be miles away before a television screen, may be listening to a playback on a tape recorder, or may be examining a "paper trail" of relevant records and documents. The lack of awareness by the examinee that he is being watched eliminates the possible reactive measurement effect that reduces the validity of direct observation. Simple observation involves focusing attention on the behavior as it occurs in a natural setting. For example, to assess the skills of a day care worker, the observer might place himself behind a one-way mirror and, using one of a large number of systems, observe and record the critical behavior. In contrived observation, on the other hand, the observer actively intervenes in a natural environment to create the necessary conditions for a specific kind of behavior to occur. For example, in order to assess altruism, one might contrive a situation such as that of a stranded motorist with a flat tire and observe whether the subject attempts to provide aid.

Performance can also be "observed" by the examination of available industrial and institutional records on individuals. Such records might indicate the amount or quality of a worker's output, the extent of lateness or absenteeism, the number of accidents, and so on.

Illustration. A student participates in a college-sponsored internship by spending a semester working with a social service agency. His work involves interviewing and attempting to respond to the needs of indigent persons in the community. The college assessor creates a situation in which several of his colleagues approach the student with contrived problems. After the interviews are completed, the "clients" record their observations according to a predetermined system and give them to the assessor for evaluation.

Development. In deciding whether to use unobtrusive observation, the assessor must determine if the student's performance is likely to be significantly affected by his awareness of the observer. If unobtrusive assessment is needed, the assessor must identify the critical elements of behavior to be observed. If the crucial behavior is not likely to occur frequently, a situation might be contrived. The observation could be conducted by one or more human observers or by special devices such as audio- and videotape. If human observers are employed, they may choose among observation systems that specify the units, categories, and methods of recording the observations. Prior to evaluation, the records must be simplified by frequency counts and other forms of data summarization.

Technical considerations. A number of problems involving reliability and validity are associated with unobtrusive observations. For example, if the observation of the student occurs in a natural setting, there is no assurance that the student's behavior is representative unless the observation occurs over a long period of time or on several occasions. Observations in natural settings and in contrived situations are not likely to be comparable from person to person unless deliberate efforts to insure comparability are made. The human observer is an imperfect data-gathering instrument. He is likely to vary in the course of his observations and he may selectively attend to only certain aspects of the subject's behavior. His limitations can be minimized by hidden monitoring equipment but ethical and legal considerations restrict the use of such devices.

A number of difficulties surround the use of institutional records in evaluating competency. The basic problem is the lack of comparability of units of output, type of work, and working conditions. For example, it is difficult to evaluate the competency of a salesman who sells 30 units of an item per week unless some standard is established for that specific work situation. Another problem in using institutional records is that they frequently yield multiple measures such as quality and quantity of output, and a way must be found to combine them into a meaningful indication of work performance.

Evaluation. The application of unobtrusive measures involving active or direct observation to the assessment of experiential learning is somewhat limited. The use of simple and contrived observation appears to have been restricted in the past primarily to classroom settings and to sociopsychological research studies. Such assessment is justified only when the performances that require assessment

are likely to be significantly influenced by the student's awareness of an observer. In most other cases, the use of such measures is prohibitively expensive.

The use of institutional and industrial records appears to be more appropriate for the assessment of experiential learning. Relatively inexpensive in comparison to the direct observation measures, the records, if evaluated appropriately, can accurately indicate the student's past accomplishments.

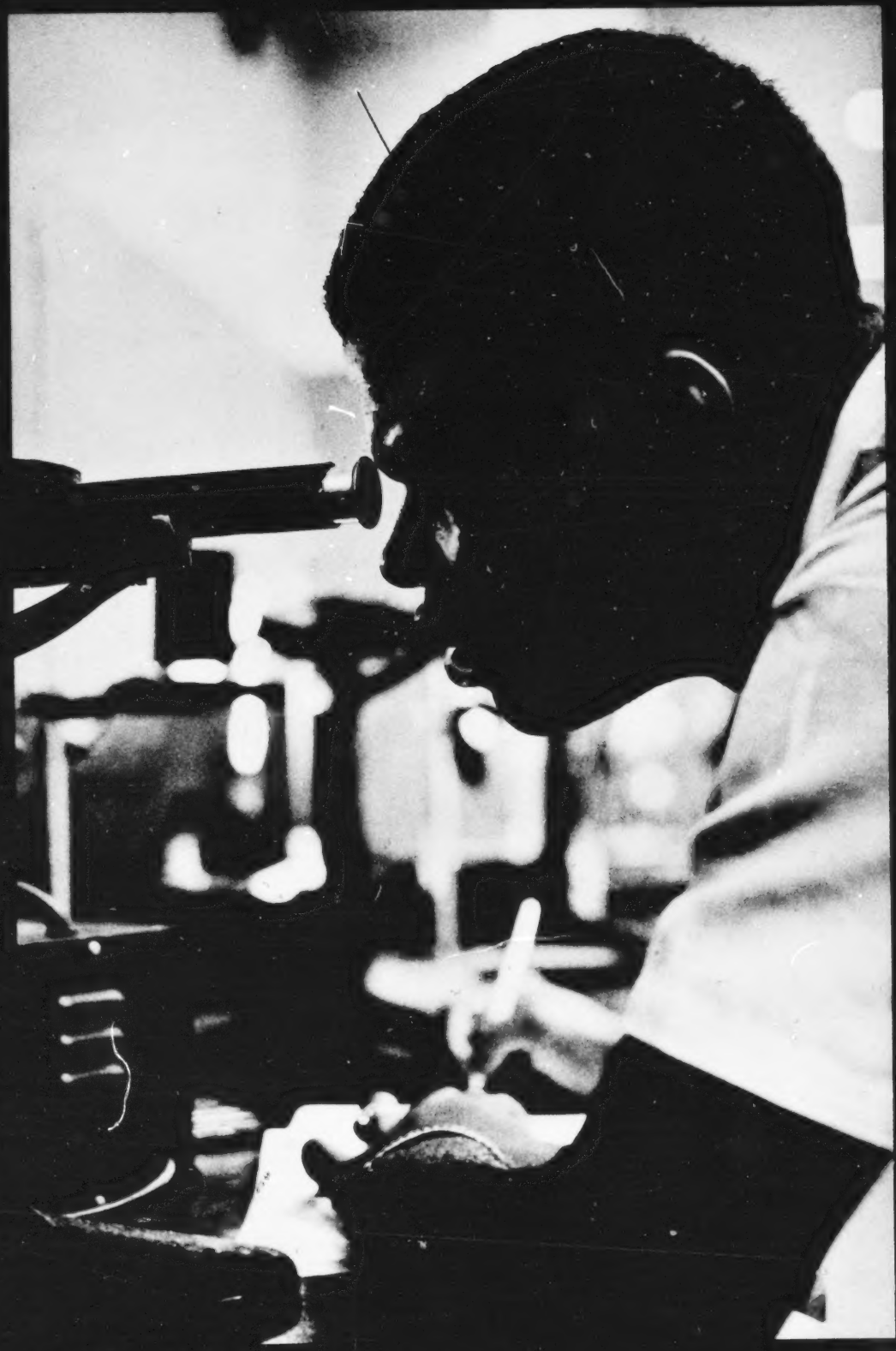
References

Simon, A., and Boyer, E.G., eds., *Mirrors for Behavior, An Anthology of Observation Instruments.* Philadelphia: Research for Better Schools, 1967.

This is a 14-volume anthology of 26 observation systems used primarily in classroom settings.

Webb, E.J., Campbell, D.T., Schwartz, R.D., and Sechrest, Q., *Unobtrusive Measures: Nonreactive Research in the Social Sciences.* Chicago: Rand McNally, 1966.

This volume describes various types of unobtrusive techniques and the reasons for using them. It summarizes a large number of investigations that employed such techniques.



III. SIMULATIONS

In simulations[1], sometimes called situational tests, the candidate is asked to pretend that he or she is engaged in some real task, the nature and content of which is described in some detail before the individual assumes his or her assigned or unassigned role. Often, the simulation is less complex and more convenient than the real situation and operates under a compressed time schedule. Such a procedure offers the assessor or assessment team the unique opportunity to observe and evaluate a performance in a lifelike setting that often cannot be assessed by other means. Unlike multiple-choice tests, essay items, and oral examinations, which typically focus on the measurement of abilities and aptitudes, simulations often concentrate on the measurement of attitudes, motivation, personality, and social skills which interact with aptitudes to get a job accomplished. For this reason, simulations are ordinarily used to measure not pure traits or skills, but instead more complex qualities. For example, they are used to measure leadership ability or decision-making skills, but not the ability to add a series of three-digit figures. Paper-and-pencil simulations include the case study technique and individual fact-finding exercises and diagnostic and troubleshooting exercises. Non-paper-and-pencil simulations include in-basket tests, management games, and leaderless group discussions.

Simulations can be thought of as having evolved from three somewhat disparate concepts: simulator trainers and simulated exercises used in the military services and space programs, gaming, and role-playing. Subsequently, simulation technology has generated a number of training and learning devices, some of which will be described later. In one annotated bibliography on simulations and academic gaming, published in 1969, the authors reported that the total literature on simulation contained more than 2,000 items if business and war games were included.

Development. For situational tests or simulations to be of maximum value in assessment, the following rules should be observed:

a) Specific behaviors relevant to the variables being measured should be identified before the setting and problem situations are developed.
b) Problem situations should be carefully structured to provide the opportunity for candidates to act in a satisfactory or unsatisfactory manner with respect to the specific behaviors.
c) The amount of fidelity to reality for each aspect of the simulation must be determined and weighed against the limits of available resources.
d) Instructions must be sufficiently clear so that the candidate knows what is expected.
e) Tasks included in the exercise should not require knowledge that the candidate is not ordinarily expected to have.

---

[1] References for all the simulation techniques described in this section are presented on pages 21-23.

f) Standard-rating instructions and multiple judgments are needed to achieve validity.

In general, the specific acts to be observed are listed on a score sheet. The assessor is required to check on the score sheet whether a participant did or did not perform a particular act that is an indicator of the identified behavior. The designer of an exercise must have a thorough knowledge of the job or work experience being simulated so that appropriate tasks can be selected to elicit the behaviors under assessment.

Technical considerations. When assessing performance in a simulated exercise, the assessor faces the problem of variability from one occasion to the next. A single observation certainly cannot be considered a reliable measure of human performance. In simulation, repeated tasks, tests, or exercises of the same type must be designed to measure the same kind of performance. Often, the greatest difficulty in the use of simulations, however, is that providing repeated measures within reasonable time limits and costs is not feasible. For example, repeating an in-basket exercise may well occupy a period of several days to a week.

Clearly, the more closely the assessment situation approximates the real-life situation, the more valid the technique; that is, one can predict the specific performances or behaviors the candidate will exhibit in a real work situation. Simulation exercises pose a dilemma in that increasing fidelity to the real-life situation is associated with increasing the validity of the measurement. On the other hand, fidelity to the actual job, task, or skill being simulated decreases the amount of control over extraneous variables and thus decreases the reliability.

Evaluation. Particular types of complex skills such as analytical thinking, goal setting, risk taking, interpersonal competence, decision making, sensitivity to the behavior of others, oral communication, and planning can be assessed using these exercises. However, the more faithful and comprehensive the simulation, the more costly the procedure. Costs involve not only money, but also factors such as personnel time, facilities, interruption of normal operations, and diversion of resources usable elsewhere.

Leaderless Group Discussion

This technique has been used specifically to identify leadership capability although it can be used to evaluate related factors such as oral communication and sensitivity toward others. Leaderless group discussions (LGD) can be of two types—with assigned and with nonassigned roles. In LGD with nonassigned roles, three to six participants are given three or four short cases or descriptions of problems usually dealing with supervisory or management situations. The participants function as consultants who must resolve the problems and submit a written recommendation for each problem. The ensuing discussion provides the group with a situation which will lead to role differentiation so that the role that the person plays is presumably similar to the role he or she is prone to adopt in natural groups. It has been suggested that the exercise chiefly measures the tendency to initiate structure in an initially unstructured situation. In LGD with

assigned roles, each participant is assigned a point of view that he or she must sell to the other members of the group.

Illustration. Several students have signed up for a practicum in political science that involves working for nominees in campaigns for state office. To evaluate whether this activity has sharpened each student's skills in goal facilitation, interpersonal competence, and oral communication, LGD with assigned roles is devised. In it, the group has acquired political party funds that it must allocate to candidates running for office. Each member of the group has a candidate to recommend for receipt of a portion of these funds, and each is given information about his or her candidate to use in order to get the largest amount of funds for his or her candidate's campaign. The group is given 90 minutes to allocate the funds, with each member "selling" his or her candidate to the others in the group. Several assessors rate each group member on three traits: prominence, goal facilitation, and sociability. Each trait is characterized by a constellation of verifiable behaviors (for example, leading the discussion, motivating others to participate, effectively saying what he or she wanted to say), which the assessor checks on a scale from "a great deal" to "not at all." Ratings are tallied for a total score for the exercise.

Development. The two major concerns in the development of the LGD exercises involve the content of the discussion problems and the delineation and description of the dimensions on which the participants must be rated. The stimulus materials must be clearly written and interesting. If the materials cannot be individualized for a specific type of experience, they should be such that the participants feel that they can enter into a discussion and appear knowledgeable. Esoteric or complicated problems will tend to intimidate the participants and result in poor "readings" on how they would act in a natural group setting.

Technical considerations. Research on LGD shows that stability over trials is high unless the time interval is long (several months) or there are radical changes in the content of the problems to be discussed in the exercise. Although industrial psychologists still are not certain what the technique measures, it appears to assess some consistent and general aspect of personality since numerous studies have shown that its scores have high predictive validity.

Evaluation. The ability to work in a group situation is difficult to judge validly via questionnaires, letters of recommendation, interviews, and paper-and-pencil tests. LGD seems to fill this assessment gap and is probably the most economical work sample of group behavior.

Management Games

Management or educational games have their origin in war games, devised in the first decade of the eighteenth century with sand simulating terrain and various objects representing troops and artillery. Games as training and assessment tools have more recently focused on skills pertaining to executive ability. Competitive encounters between individuals that involve some degree of skill or luck, games

operate, in the educational sense, chiefly as a function of problem-solving skill. Games as simulation exercises can be categorized as 1) media-ascendent simulation games mediated by machines; 2) interpersonal ascendent simulation games in which decision making, role playing, and player interaction are emphasized; and 3) nonsimulation games which merely provide a competitive context for learning concepts and principles.

Illustration. A candidate is enrolled in a special cooperative education program in business and management. Over a two-year period, she alternates course work and work experience seminars with work in every department of a small manufacturing firm. Several instruments are used to assess students in this program, one of which is a business game. The candidate and three other students in the program participate in a game in which component parts are bought for the manufacture of prototype products. The team is instructed to maximize profits, which involves deciding what parts and how many to buy and what products to manufacture, depending on prices being offered. The student, with the other players, is rated by business faculty on her planning and organizing skills, leadership behavior, communication and problem analysis skills, judgment, initiative, decisiveness, and flexibility.

Development. The first step in game design is to outline the scope and detail of the game problem being considered in terms of geography, time duration, actors concerned, types of actions or functions involved, and the objectives of the actors. The next step is to determine the adversaries or competitors. These may be individuals who occupy specific roles, groups, companies, communities, or entire nations depending on the scope and context of the problem. Each actor must be distinguished from the other players by the decisions he makes, resources available, and preferences held. Resources made available to the actors to solve the problem must be identified and defined in such terms as money, time, and energy. When criteria are established, they are usually expressed in terms of achieving certain specific objectives with a minimum expenditure of resources or a maximum number of objectives by using a limited number or amount of resources. Finally, the sequence of tasks needed to get the actors from start position to the terminal position in the situation is determined.

Technical considerations. Little research or discussion has been focused on measurement problems concerning the use of games for assessment purposes. Several factors inherent in the technique might have a tendency to confound outcomes and thus interfere with reliable and valid measurement. For example, performance in gaming can be influenced by variables such as tolerance for ambiguity, risk taking, and locus of control. These in turn are moderated by sex, minority group membership, and socioeconomic class. If these variables influence problem-solving behavior and decision making in the world of work, the use of games to measure these skills is justified. However, if factors such as risk taking only influence behavior in the gaming situation, this sample of the candidate's behavior would not be an accurate representation. Also, some individuals who participate in a game as an assessment technique may resent being judged by winning or losing. No clear-cut research shows that gaming skill is related to job skills or paper-and-pencil tests. On the other hand, there is

evidence of a negative relationship between grades in business courses and game skills. This leads one to hypothesize that skills or dimensions tapped by gaming behavior cannot be measured entirely through course grades or test scores.

Evaluation. Academic or management games could be one of the most economical assessment techniques in that virtually hundreds of them are available through companies that specialize in manufacturing educational products. There is no evidence that "canned" games make a difference in eliciting the desired behaviors; therefore, it is probably not necessary to develop tailor-made games for each assessment situation. Since the scoring and rating behavior during an educational game is usually complex, the games shouldn't be used as the sole measurement of problem-solving or decision-making ability.

Interview Simulation and/or Role Playing[2]

One form of situational exercise may employ role playing to capture a sample of interpersonal, job-related behavior. The situations set up for role playing involve the kind of person-to-person behavior characteristic of work in education, business, administration, and human services. The most common role-playing script involves a simulated interview in which the assessee plays a role such as personnel manager, supervisor, or customer service representative.

Illustration. A student seeks credit for work experience she obtained in a property management firm. Her main duty in that job was to intercept phone calls dealing with problems that lessees were having with various properties listed with the firm. She participates in a telephone interview simulation in which a client has several complaints concerning his apartment. Her performance is videotaped, and trained evaluators judge her ability to communicate in a stressful situation, impact, listening skills, tenacity, and flexibility.

Stress Interview

In the stress interview, the candidate is deliberately placed in a condition of stress without his or her knowledge. To induce stress, the candidate may be interrupted or criticized for his or her opinions, and the interviewer may keep silent for an extended period. Under these conditions, the candidate can be evaluated for tendencies toward a lack of emotional control. If certain problem-solving questions are asked during such a interview, an opportunity is provided to evaluate how well the candidate can think or function under stress.

Illustration. A student requests credit for prior experience as a community organizer. He is presently enrolled in a program leading to a degree in community planning. His mentor lists several skills and learning outcomes that he

---

[2]When the development, technical considerations, and evaluation for a particular type of simulation are essentially the same as for simulations in general, these topics do not receive separate coverage.

feels are relevant to the student's degree and that should have been acquired from his job experience. One of these is the ability to cope with stressful situations. A stress interview, containing problem-solving questions, is conducted by the mentor and videotaped. Several members of the social science faculty view the videotape and rate the student on such factors as emotional control, tact, sense of humor, and decision making under pressure.

Technical considerations. The ability to withstand verbal stress during an interview may not be related to the ability to withstand the types of stress found in many types of work. Available validity data on situational stress tests are meager. One problem is the unreliability of criterion ratings, which tends to reduce validity estimates. For example, stress ratings for the candidate on the job are less reliable than those ratings he might receive in a controlled assessment situation. Any assessment interview is a sufficiently stressful situation, and behavior during the initial stages of the interview can be a valid indication of reaction to stress.

Evaluation. Stress interviewing is highly dangerous and should never be used on candidates who, in nonstressful situations, show signs of emotional disturbance. It may be well to schedule two interviews. The first should be conducted in a normal, relaxed manner as a tryout to see if the candidate shows signs of extreme nervousness. If the candidate reacts well, the second procedure, the stress interview, should be scheduled. The most appropriate use for the stress interview in assessment would be to evaluate behaviors where reaction to stress is extremely crucial, such as ambulance driving, police work, and ombudsmanship. For the everyday pressures inherent in any job, stress reactions probably are not as crucial to job performance. If such a rating or evaluation is desired, it could be obtained through simulation techniques which emphasize decision-making ability under time or group pressure.

Written Simulations

A written simulation is appropriate when the candidate population is large, the expense of simulation is too great, or the assessor is interested in obtaining data from the candidate on specific crucial information and skills that he or she must have to solve a problem or diagnose human or machine malfunction. A paper-and-pencil situation, it may follow any number of written formats, such as multiple-choice or essay.

Diagnostic and trouble-shooting exercises. The most widely used forms of this type of exercise are concerned with the measurement of diagnostic skills necessary for repairing various types of instruments and machinery. Each item on the test briefly describes a specific malfunction which is followed by a list of checks the repairman might wish to make in order to diagnose the problem. Beside each check is an indication of the result of that check covered by a chemical coating. When the examinee selects a particular check or procedure, he scratches the coating and the results of the check become visible. For example, an individual taking a television repair test would read the symptom—"No picture, screen

dark." Ten check procedures are listed under the symptom. He selects the check that says "Turn up horizontal and vertical size control." and scratches the box alongside this point, which says "No effect." He must then decide which checkpoint to select next in order to get more information about the problem. Eventually, he must make a diagnosis. The examinee's score is diminished if he selects more checks than are necessary to diagnose the problem. Since this is a linear approach to problem solving, a premium is necessarily placed on efficiency in reaching a single correct decision or on the appropriateness of each checkpoint decision considered independently.

Recently medical education has devised a much more complex version of this technique to measure the diagnostic skills of practicing physicians and medical students. The main difference between these patient management simulations and the technique described above is that the medical exercises consist of branched problems and require the examinee to make choices from an almost unlimited number of broad strategic routes, more than one of which may lead to acceptable results.

Case Study Method

This paper-and-pencil simulation involves a fairly complete and realistic statement of some problem that can be related to business, human services, or almost any area of human experience. The candidate is asked to write an analysis of the problem in a specified period of time, and his or her report should contain alternative solutions and recommendations for action. The exercise is used to assess the student's ability to analyze and solve complex problems. Sometimes, the case under consideration is not written. It can be presented to the candidate in several modes, such as filmstrip, tape, or videotape.

Illustration. A candidate contracts to work as an aide in a drug rehabilitation center. To show his mentor that the objectives of the learning experience have been accomplished, he is required to do a case study exercise. The mentor has a set of taped cases or situations reported to him by a senior drug counselor. They consist of poor approaches to handling counseling situations. The candidate is asked to analyze several situations by explaining the errors of commission and omission on the part of the counselor and indicate how he would have handled the client if he had been counselor. The cases have been preanalyzed by experts in drug rehabilitation, and their analyses are used by the mentor as a standard by which he can judge the candidate's responses.

Individual Fact-Finding and Decision Making

This type of exercise is not a pure paper-and-pencil simulation measure but rather a combination of the interview technique, oral or written examination, and the case study approach. The candidate is given a specific time limit within which he or she must prepare a line of questioning leading to the solution of a problem that has been described. Next, the candidate interrogates a person in a resource role for a specified time limit. At the end of the time period, the

candidate must arrive at a decision concerning the problem and must reveal this decision and the reasoning behind it in an oral or written report. The resource person may then question the participant to test the quality of his reasoning, his persuasiveness, his ability to think and talk under pressure, and his grasp of the issues underlying the problem. The resource person may evaluate the candidate, or a panel may observe the whole exercise via videotape and make judgments concerning the adequacy of the candidate's responses.

Illustration. A candidate applies for credit for prior experience as personnel manager of a small department store, and the assessor presents the following problem to him. The lingerie department head in a clothing store claims that one of the salespersons in his department has taken home store items in a bag on several occasions, but the employee denies the accusation. The assessor, the person in the resource role, is questioned by the examinee for further information. After the examinee has written a report or prepared an oral presentation concerning the resolution of this problem, an evaluation team, which includes the assessor, questions him in order to obtain more insight into the reasoning that led to his decision. He is then rated according to a checklist of decision-making skills by the evaluation team.

In-Basket Test

The in-basket exercise takes its name simply from the in-basket or tray found in most businesses and institutions where employees work in an office or at a desk. Originally, the exercise was designed for individuals being assessed for managerial capability in business, but variants of the test are now used to assess policemen, military officers, and school principals. The exercise simulates a specific organizational role, requires the examinee to deal with problems as if he or she were actually on the job, and provides a set of materials that may contain incoming letters, memos, reports, papers to be signed, and telephone messages. These documents constitute the test items. Prior to the candidate's performance, he or she has the opportunity to study background materials for orientation in the role to be assumed for the hypothetical job. During the exercise, the candidate is instructed to handle all the documents in the in-basket as would be done on the job. All candidate decisions and actions are recorded in writing, which may include letters, memos, outlines, directions, and the content of phone calls. These written products are usually scored by trained evaluators using methods similar to those used for essay grading. More efficient scoring techniques have been devised wherein a checklist is provided to the candidate and he or she picks out the actions performed from a wide list of possible actions. The candidate may also add any unusual or unlisted actions taken. These responses are judged and scored by trained assessors. At times, an interview is used in lieu of scoring procedures.

Illustration. A student enrolled in a restaurant management program wishes to receive credit for her experience as a school cafeteria manager. A modified one-hour in-basket exercise is devised that contains various authentic-looking documents related to problems in purchasing, food preparation, menu planning,

personnel allocation, and budgeting. The in-basket is not scored. Instead, the student's academic adviser interviews her about her approach to the tasks and reasons for taking the actions indicated. The adviser writes a report of the interview with a recommendation for or against credit and gives it to a faculty member for another judgment.

Development. A typical series of steps in the development of an in-basket test might include the following:

1) Conduct a task analysis of the job to be simulated.
2) Duplicate all paper work related to the job for a week.
3) Interview a number of individuals who are highly skilled at the job under consideration.
4) Develop a pool of sample documents to be reviewed by people on the job.
5) Administer an experimental version to individuals on the job.
6) Obtain ratings of the individuals from supervisors.
7) Analyze data and prepare a final version on the basis of these data.

Technical considerations. A candidate's behavior on one in-basket test usually cannot be replicated on a second independent in-basket test; therefore, it is difficult to determine the stability of such measures. Also, in view of the constraints on time and information in the artificial test situation, it is possible that the style that the candidate displays during the test is not representative of the style he will display in actual performance on the job.

Evaluation. Although the in-basket test is expensive to give and score, it is an appraisal approach that has a great deal of face validity, so that it appears to be a fair and realistic sample of a job. Because of problems concerning its reliability and predictive validity, it is being used more and more for training rather than assessment.

References on Simulations

Abt, C.C., *Serious Games.* New York: The Viking Press, 1970.

This book explores the ways in which games can be used to instruct and inform. Chapters examine topics such as evaluating the cost-effectiveness of games and improving education with them. The use of games in general areas such as the physical and social sciences, vocational guidance, and government and industrial planning is discussed.

Barton R.B., *A Primer on Simulation and Gaming.* Englewood Cliffs, NJ: Prentice-Hall, 1970.

This book combines computer and noncomputer approaches to simulations and gaming but does not require learning a computer language. A checklist for conducting man-only and man-computer simulations is of practical value.

Bass, B.M., "Analysis of the leaderless group discussion." *Journal of Applied Psychology*, 1949, 33, 527-533.

One of the first attempts to evaluate the LGD technique quantitatively, this article focuses on the investigation of the amount of agreement among raters of discussion participants and the relationship between the amount of time the participant talked and the ratings he obtained.

Bass, B.M.,"The leaderless group discussion as a leadership evaluation instrument." *Personnel Psychology*, 1954, 7, 470-477.

The article summarizes the results of LGD research.

Brody, W., "Judging candidates by observing them in unsupervised group discussion." *Personnel Journal*, 1947, 26, 170-173.

This is a "how to" article that guides the reader in LGD development and administration.

Crooks, L.A., "Issues in the development and validation of in-basket exercises for specific objectives." *Research Memorandum* (RM-68-23), Princeton, NJ: Educational Testing Service, 1968.

This paper discusses the practical issues surrounding the development and use of the in-basket test.

Fiske, D.W., "Why do we use situational performance tests?" *Personnel Psychology*, 1954, 7, 464-469.

The article discusses sources of error in measurement when using situational tests.

Flanagan, J.C., "Some considerations in the development of situational tests." *Personnel Psychology*, 1954, 7, 461-464.

The author explains the advantages and disadvantages of using situational tests to measure attitude, motivation, and other personality factors.

Frederiksen, N., "Factors in in-basket performance." *Psychological Monographs*, 1962, 76, (Whole No. 541).

A factor analysis of the 70 aspects of behavior used to rate 335 people on an in-basket test revealed 8 primary factors, including informality, directing subordinates, discussing, and acting in compliance with suggestions.

Green, T.B., "Do filmed cases improve the case method?" *Training and Development Journal*, 1973, 27, 28-31.

The article summarizes a study in which the case method was combined with a film approach.

Lopez, F.M., "A management tool: The in-basket exercise." *Simulation and Gaming News*, 1973, 1, 1 and 19.

The article describes the Quasar In-Basket Exercise which can be adapted for use in almost any organization. It includes a method of scoring via the Action Report, either by mechanical or electronic means.

McGuire, C.H., and Babbot, D., "Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement*, 1967, 4, 1-9.

This article describes the development of a series of branched problems in patient management that require sequential analysis and decision. The problems were designed to measure aspects of behavior defined by a criterion group as essential components of clinical competence.

Parry, S.B., "The name of the game is simulation." *Training and Development Journal*, 1971, 25, 28-32.

The writer explains the rationale for using games and simulations, with some examples, and makes the interesting distinction between simulations designed as application of prior learning and those intended to give the learner an opportunity to gain insight into his own behavior.

Spivack, J.D. "Critical incidents in counseling: Simulated video experiences for training counselors." *Counselor Education and Supervision*, 1973, 12, 263-270.

This article explains the rationale, development, and use of a simulation approach to training counselors in which videotaped vignettes of critical incidents that occur in the course of establishing and maintaining a helping relationship are presented to trainees as stimuli.

Tansey, P.J., and Unwin, D. "Sources in simulation and academic gaming: An annotated bibliography." *British Journal of Educational Studies*, 1969, 17, 193-208.

This bibliography attempts to reduce more than 2,000 articles on simulation to 87 citations. The articles selected concentrate in the fields of school and college learning with particular emphasis on simulation as a teacher training tool. Some articles on role playing and sociodrama are included.

Williamson, J.W., "Assessing clinical judgment," *Journal of Medical Education*, 1965, 40, 180-187.

The author describes methods of analysis for data generated from branched problems in patient management.

IV. ASSESSMENT CENTERS

The first known applications of the assessment center method occurred during World War II in England for the purpose of selecting army officers. Subsequently, the United States adapted the method for the selection of intelligence officers for the Office of Strategic Services. In recent years, the approach has been reconceptualized and used mainly by industry to identify managerial and administrative talent, although various aspects of the approach can be adapted for assessing nearly every type of aptitude and skill relating to job performance.

As it has been known in industry, the assessment center is a comprehensive, standardized procedure in which multiple appraisal techniques are used in combination to evaluate individuals for various purposes. An assessment center usually functions in a secluded atmosphere with individual groups consisting of 12 assessees. The multiple assessment procedures may be of a week's duration, although firms such as American Telephone & Telegraph have developed one-day assessment centers to accommodate a high volume of candidates. The appraisal techniques generally used fall into three categories: personality tests, simulation exercises, and paper-and-pencil measures. The following procedures might be included in an assessment center: interviews; an in-basket exercise; leaderless group discussion; a management game; projective tests such as the Rotter Incomplete Sentences Blank; paper-and-pencil tests and questionnaires such as the Edwards Personal Preference Schedule, the School and College Ability Test, and a personal history questionnaire; a case study problem-solving exercise; and interview simulations. Most of these techniques have been described and explained elsewhere in this booklet. In addition to the individual procedures for evaluating candidates on specific exercises, the assessment center staff, usually a team of six to nine individuals, assembles, reviews, and discusses all of the results immediately following the completion of the assessment center exercises. Each candidate is evaluated separately, a process that takes one to two hours. A typical evaluation consists of reading the assessee's autobiographical information and interview transcript and evaluation. Reports on the simulated exercises, paper-and-pencil instrument scores, and projective test analyses are next integrated by the assessors. Then each member of the assessment team independently rates each candidate on a list of characteristics or on-the-job criteria, which may include organizing ability, planning, decision making, problem solving, creativity, communications skills, and sensitivity toward others.

Although assessment centers have been conducted primarily in industry, the centers can serve an important function in institutions of higher education making a concentrated effort to better integrate assessment with instruction and guidance. The center serves to consolidate the assessment activities within a college and focuses its resources on the measurement and evaluation of the many different facets of students' learning. While a center might conduct those assessment activities that most faculty are not trained to perform, it might also provide other services, such as those of a faculty resource center, monitoring and improving assessment procedures, and training mentors in the assessment process.

Illustration. Ten business majors are participating in a work experience program in which they have been assigned to internships as salespersons in various business firms. At the conclusion of the internship, the students request credit for their experience. A one-day assessment center is scheduled for them in which they participate in procedures that include a leaderless group discussion involving a sales promotional problem, a case study that describes a problem concerning a drop in sales of a previously fast-selling product, and an interview that focuses on ascertaining the various tasks accomplished and skills learned on the job. Each participant is evaluated individually by the assessment team as well as by each assessor. Criteria used are the projected learning outcomes set forth in the internship contract. The assessment team is made up of faculty from various areas in the business and management sciences.

Development. A first step in planning an industrial assessment center involves devising a list of the qualities relevant to success in the job or job area for which the candidate is being assessed. In order to arrive at this list, an industrial psychologist or job analyst should interview various personnel actually working successfully in the occupation and related positions for which the center is planned. A preliminary list of qualities or dimensions is selected from "canned" materials or is tailor-made for specific problems in unique settings. A group of trained assessors under the direction of an industrial psychologist applies these procedures to a trial group considered successful in the field under consideration. The results of the pretest should indicate any redefinition of qualities or appraisal techniques that would make the center successfully operational.

Developing an assessment center in a college is likely to involve somewhat different considerations than those in establishing an industrial center. Among the considerations are the types of learning experiences to be assessed, the techniques to be used, and the way in which the faculty and students will be involved.

Technical considerations. Technical research on assessment centers has focused on studies of the interrelationships between the rating variables and overall predictions of the assessment staff, studies of the assessment techniques with particular reference to their contributions to the assessment process, and studies of relationships of staff evaluations to behavioral criteria such as survival and progress in the world of work. This research can be summarized briefly in the following statements:

1) It is difficult to find relevant, reliable, and sensitive criteria of job success in many work situations.
2) Criterion variance is difficult to obtain because there are few poorer performing assessees inasmuch as only the better performers are promoted or even hired.
3) In some experimental-control studies, results have shown a positive relationship between assessment performance and management success.
4) Studies to determine the contribution of individual techniques reveal that an overall, global assessment rating given to assessees by the assessment team is a better predictor of job success than any single measurement component.

5) Situational techniques increase the criterion variance accounted for when compared to paper-and-pencil measures such as objective tests.
6) Assessment results correlate more highly with on-the-job ratings of the candidate done by special observational teams than with ratings done by on-the-job supervisors.

It should be noted that the above research has viewed the assessment center as a selection and promotional tool, with predictive validity emphasized. When the assessment center approach is used to evaluate a student's work experience in order to award credit for prior or institutional sponsored learning, concurrent or convergent validity is technically appropriate. That is, assessment center data should be correlated with on-the-job performance measures such as supervisor's ratings, peer- and self-ratings, and employee's documentation. For this type of appraisal, learning outcomes are the target of assessment rather than the potential for learning or acquiring job-related skills.

Evaluation. The assessment center technique is an expensive approach to the evaluation of learning in terms of development and subsequent staff time for service on assessment teams. However, if a large number of candidates is enrolled in a limited number of work experience programs, it may be the most efficient form of assessment. Of course, the assessment center procedures could be limited to a smaller number of techniques than those used in elaborate centers that concern themselves with management programs. There is also the possibility of using assessment packet materials prepared by various consulting firms, which are available for a nominal fee per candidate. Workshop sessions still are needed to train faculty and counselors for their roles on the assessment team. Recently this approach has been adapted for use in colleges that emphasize competency-based instruction and work experience. The center approach offers candidates the unique experience of self-awareness and self-evaluation through assessment.

References

Bray, D.W., and Campbell, R.J., "Selection of salesmen by means of an assessment center." *Journal of Applied Psychology*, 1968, 52, 30-41.

This article describes how an assessment center was devised for candidates for sales positions. Appraisal techniques consisted of paper-and-pencil tests, an interview, and individual and group simulations. Assessment results were compared with job performance some months later as evaluated by a special observation team.

Bray, D.W., and Grant, D.L., "The assessment center in the measurement of potential for business management." *Psychological Monographs*, 1966, 80, (Whole No. 625).

The assessment center process in the Bell System's Management Progress Study is described. The results, based on 355 young managers, show that of the techniques studied the situational methods and paper-and-pencil ability tests are more predictive of job progress than personality questionnaires.

Bray, D.W., and Moses, J.L., "Personnel selection." *Annual Review of Psychology*, 1972, 23, 545-576.

A section of the paper supports the authors' claim that the use of broadband assessment methods such as the assessment center will continue to grow and that, although they will be used primarily with management personnel, their possible applicability to rank-and-file employees, blue-collar workers, and college students will be investigated.

Crooks, L.A., "The selection and development of performance measures for assessment center programs." *Research Memorandum* (RM-74-6) Princeton, NJ: Educational Testing Service, 1974.

This paper contains a brief description of almost every technique in use in assessment centers. The author explains how a set of these techniques can be integrated for a senior-level assessment program and an early identification of talent program.

Office of Strategic Services Assessment Staff, *Assessment of Men.* New York: Rinehart, 1948.

This book gives a complete accounting and history of the development of assessment centers for the Office of Strategic Services during World War II. Each technique used is described in detail and the evaluation of the assessment procedures themselves is summarized.

V. ESSAY EXAMINATIONS

Essay examinations are instruments characterized by the presentation of questions or tasks to which the student responds by organizing and writing an answer. The response required may be brief or involve an extensive exposition. Some tests require the student to give written responses to situations presented on audio- or videotape. Usually, no single response or pattern of responses can be listed as correct, in that even an expert cannot usually classify a response as categorically right or wrong; rather, different degrees of quality or merit can be recognized. A structured essay is one in which a topic is assigned with a number of words expected or with other conditions set. An unstructured essay allows the candidate to write on any relevant subject, in any form or length he or she chooses. Essay content for either case may be factual or creative.

Illustration. A candidate applies for credit for creative writing experience that she acquired through nonfiction free-lance writing assignments for women's magazines. The candidate is presented with a representative list of topics and is required to select one topic and write a short piece of nonfiction on it. The essay is evaluated by several judges along dimensions such as imaginativeness, accurate vocabulary, and structure.

Development. The following steps should be considered in constructing essay questions:

1) Define objectives clearly so that tasks may be constructed which are related to the framework of objectives.
2) Formulate well-focused questions and define the task through determining the amount of freedom of response.
3) Construct model answers while taking note of writing time.
4) As a pretest check, have an examiner who didn't formulate the questions answer them.
5) Ask colleagues or specialists to review the questions.
6) Pretest the items so that the following can be answered:
   a) Do students appear to understand the intent of the question, or do they appear to interpret it in ways not intended?
   b) Is the question of appropriate difficulty for the examinee population?
   c) Is it possible to grade answers reliably?
7) Recognize the importance of an adequate sample of questions (a sufficient number, the right number of short and long questions, and alternate questions from which to choose).

Technical considerations. In any discussion of the reliability of essay assessments or examinations, several sources of error must be considered, such as interrater reliability and variability of examinee performance from one sample of questions to another. The amount of time given to writing the essay will influence the consistency of the performance, which will vary

from student to student. However, the main source of error is associated with the grading of essays. Essay rater research has shown that:

1) different raters tend to assign different grades to the same paper,
2) a single rater tends to assign different grades to the same paper on different occasions, and
3) the differences tend to increase as the essay question permits greater freedom of response.

To the extent that an essay assessment measures a factor or set of factors which are unique and cannot be measured by the use of other assessment techniques such as objective exams, inventories, and management games, the utility of the exercise is assured. This can be shown by correlating essay ratings with scores on other types of assessment instruments. If the correlation between the essay and another instrument is high, then both procedures essentially measure the same ability, skill, or factor. Clearly, if the skill to be measured is the ability to communicate effectively in the written mode, the essay test is the most valid procedure. If, however, the candidate is to display his or her knowledge about a certain historical topic, a multiple-choice test procedure probably would be more economical and valid.

Evaluation. Essays should be used when attempting to assess situations that involve demonstrations of the following abilities or competencies:

1) presenting knowledge of areas in a written format, including such products as reports and memos,
2) expressing thoughts effectively on some topic of wide human interest,
3) translating into English a passage presented in another language,
4) outlining a research design in the hard or soft sciences,
5) writing creatively,
6) simply communicating effectively in the written mode.

Essay assessment may be used when other forms of assessment are prohibitively expensive or impractical. For example, if a student has contracted for a cross-cultural experience in Mexico, reports or essay items concerning the student's work may be one feasible choice of technique for ongoing assessment in view of his inaccessibility.

References

Coffman, W.E., "Essay Examinations," in Thorndike, R.L., ed. *Educational Measurement*. Washington, DC: American Council on Education, 1971, 271-302.

This chapter gives a systematic analysis of essay examinations and describes their construction, limitations, scoring procedures, technical problems, and needed research in the area.

Ebel, R.L., *Measuring Educational Achievement*. Englewood Cliffs, NJ: Prentice-Hall, 1965, 84-123.

Chapter 4, "The Characteristics and Uses of Essay Tests," is of value to the classroom teacher. Special consideration is given to guidelines for constructing teacher-made essay tests.

French, J.W., "Schools of thought in judging excellence of English themes," in *Proceedings, Invitational Conference on Testing Problems, 1961*. Princeton, NJ: Educational Testing Service, 1962, 19-28.

This paper investigates the reasons why essay graders disagree and what information the essay response, when properly graded, can offer that the multiple-choice response cannot.

VI. OBJECTIVE WRITTEN EXAMINATIONS

This appraisal technique has been used so often and so long in education that information concerning its advantages, disadvantages, and technical considerations (such as reliability and validity) can be obtained from the references listed here or from any basic text in educational measurement. It will probably be most useful here to discuss briefly types of items other than the well-known multiple-choice item that can be helpful in measuring learning outcomes.

The completion form. This item form is characterized by the presence of a blank in which the examinee writes the kind of answer called for. It has fallen into disuse in recent years mainly due to technical problems concerning keying and scoring. However, the ability to produce the correct answer for a short-answer item probably represents a higher level of competence that does the ability to point out which of the choices in a multiple-choice item is best or correct. Despite the seeming superiority of this item type in measuring competence, it is not used extensively because it is so difficult to write. The item must be written so that the examinee knows exactly what response is expected. Even so, there usually are several ways the correct answer may be expressed. As the creditable responses to a completion item increase, the scoring key becomes increasingly cumbersome and the obtained score less reliable. Because of these problems, the following pointers should be considered when writing completion-form items:

1) Use the completion form only for questions that can be answered by a unique word, phrase, number, or symbol.
2) Do not borrow statements verbatim from context and attempt to use them as completion items.
3) Make the question or the directions explicit.
4) Allow sufficient space for students' answers and arrange the spaces for convenience in scoring.
5) In computational problems, specify the degree of precision expected.
6) Avoid the use of too many blanks in an item.

The true/false form. In the early days of objective testing, the true/false item was as popular as the multiple-choice item is today; however, it is probably the most difficult type of item to prepare. There are few statements in any subject area about which one can say they are universally true or false. Therefore, it is inappropriate to use this item form to tap thought processes such as problem solving, analysis, and interpretation. True/false items tend to test lower mental processes such as rote memory; thus the information they measure is usually trivial. Item analyses have shown that often the brightest and most creative individuals in a group make the lowest scores on true/false tests because they always can think of an exception to the statement. Also, even the poorest student guessing haphazardly will have a 50-50 chance of answering each item on a true/false test correctly. One major advantage of true/false items is that if there

are time and space constraints in paper-and-pencil testing, the items permit a larger sampling of a subject-matter area. The following pointers should be considered when writing true/false items:

1) Base true/false items only on statements that are true or false without qualifications.
2) Avoid the use of long, involved statements.
3) Avoid the use of statements lifted out of context.

The matching form. The matching test item consists of a list of premises, a list of responses, and a set of directions for matching one of the responses to each of the premises. Both responses and premises must be relatively homogeneous. Sophisticated variants of the matching form are particularly well-suited to item topics dealing with explanations, criticisms, and other higher-level learning products. In writing matching items, one should observe the following rules:

1) Group only homogeneous premises and homogeneous responses in a single matching item.
2) Use a relatively short list of responses.
3) Arrange premises and responses for maximum clarity and convenience to the examinee.
4) In the directions, explain clearly the basis for matching.
5) Do not attempt to provide perfect one-to-one matching.

References

Ebel, R.L., *Measuring Educational Achievement*. Englewood Cliffs, NJ: Prentice-Hall, 1965.

This volume deals primarily with achievement tests prepared by teachers and professors for use in their own classes, and it emphasizes methods of test development and item analysis rather than the selection and use of standardized tests. It includes practical suggestions for planning, constructing, administering, and scoring classroom tests, and for analyzing the results. No previous special training in educational measurement is assumed.

Sanders, N.M., *Classroom Questions: What Kinds?* New York: Harper and Row, 1966.

This volume is intended for teachers and other makers of classroom test questions and is based on Bloom's *Taxonomy of Educational Objectives*. Chapters include Questions Designed for More Than Memory, Memory, Translation, Interpretation, Application, Analysis, Synthesis, Evaluation, and Planning for Questioning.

Wesman, A.G., "Writing the test item," in Thorndike, R.L., ed. *Educational Measurement*. Washington, DC: American Council on Education, 1971.

This chapter points out the problems in writing items for objective tests and indicates the situations for which various item types can or cannot be used.

Wood, D.A., *Test Construction: Development and Interpretation of Achievement Tests*. Columbus, OH: Charles E. Merrill, 1960.

Addressed primarily to teachers, the book's 10 chapters discuss principles and procedures of test construction, with emphasis on objective achievement tests. There is some coverage of essay testing and the construction of aptitude tests.

VII. INTERVIEWS

Often defined as a conversation with a purpose, the interview has been the main method of collecting social research data and of course is a pivotal procedure in the personnel selection process. The interview is a face-to-face, interpersonal role situation in which one person, the interviewer, asks the candidate questions designed to obtain answers which will lead to an evaluation of one or more learning outcomes from life experience and/or training. During the interview, skills, knowledge, or concepts attained through experience can be listed, described, and later evaluated. The assessment interview could be composed of items that represent critical incidents related to the job or experience under consideration.

Development. When planning and constructing the assessment interview, one should be aware of several major pitfalls, such as the multiple meaning and ambiguity of words; a lack of sharp focus on the skills, aptitudes, and experiences under assessment; and a lack of appreciation of the interview as a measurement instrument.

There are three main types of interview questions:

1) *Closed questions* offer the candidate a choice among two or more alternatives such as yes-no or agree-disagree.
2) *Open-ended questions* supply a frame of reference for the answers but do not constrain a candidate with respect to the expression of his or her answers.
3) *Scale items* constitute sets of verbal statements whereby the candidate responds by expressing degrees of agreement and disagreement.

The closed question is appropriate when the objective is limited to the classification of the interviewee with respect to some attitude, perception, or other clearly understood dimension. Such questions have greater reliability, but the information gained from them tends to be superficial and they have a tendency to frustrate the interviewee when he or she finds none of the alternatives suitable. Open-ended questions permit the assessor to probe in order to ascertain possession or lack of knowledge on the part of the assessee and to establish his or her frame of reference. Scale items can be used as a validity check against open-ended information.

In general, questions in the interview instrument should:

1) be directly related to the skills and knowledge the assessor wishes to assess,
2) be appropriate in type for the kinds of information to be obtained,
3) be clear and unambiguous,
4) not include leading questions,
5) not demand knowledge or information that the candidate does not have or should not be expected to have,
6) not require personal information that the interviewee may resist revealing,
7) not be slanted so that the interviewee offers the socially acceptable response.

Technical considerations. Reliability of the interview as a measurement instrument involves the accuracy of the information provided by the interviewee and the reliability of recording and evaluating the data by the interviewer. Studies revealing that meaningful and reliable data can be reported from the assessment interview separate the interview from an evaluation. For example, ratings of interview variables and dimensions are made by coders from the interviewer's transcribed summary, not by the interviewer. The accuracy of information offered by interviewees has been shown to vary due to such factors as repression and memory decay.

Researchers have long suspected that institutional records, information, application blanks, resumes, grades, and other documentation concerning a life or work experience are less credible and valid than the "subjective" account given by the candidate in an assessment interview. For example, in one validity study interviewers' ratings on competences or dimensions in a subject area were compared with faculty and experts' ratings on the same skills, using a recording of the interview. Correlations between the expert ratings and interviewer ratings were higher than correlations between the expert ratings and grades in the subject area that contained the competence. However, even though some studies indicate that the interview produces valid judgments about the interviewer, it would seem that convergent validity studies should be undertaken whereby the interview information is compared with other measures of competence (for example, employee ratings, anecdotal reports from supervisors, documentation in the form of references, checklists, and attitudinal scales). An extreme example of invalid measurement would be that in which a candidate lists his duties on the job and rates himself high on the accomplishment of job tasks—but an interviewer finds later that his performance and punctuality are so poor that the candidate is asked to resign from his position.

Sources of interviewer error in evaluating the interviewee and the information he or she offers include all those in any judging or rating situation, such as the halo effect, overgeneralization, and a tendency to leniency (see the section on ratings).

Evaluation. The interviewing technique in assessment has many important qualities that objective tests, scales, and observation instruments do not possess. It is flexible and adaptable to individual situations, and, in most cases, the interview allows for probing in order to obtain more information. A major disadvantage is that it takes much staff time. If many candidates must be assessed, more economical methods should be employed or a particular type of interview method should be used (see below). However, when information is not a matter of record or readily obtainable from sources other than the candidate himself, some type of interview is indicated. Several types of interviews are described below.

Structured-Unstructured Interviews

In the standardized or structured interview, the questions, their sequence, and wording are fixed. The interviewer can take very little liberty in changing the interview format, which is composed mainly of closed questions. The highly structured interview is sometimes considered synonymous with the questionnaire and can be administered by tape, in person, or in a written form. However, the responses to such an instrument are oral.

The unstructured interview is much more flexible and open. Topics, issues, and questions are outlined but the content, sequence, and wording of questions are entirely up to the interviewer. The questions in such an interview are usually open-ended and unstandardized as to stimulus and response categories. In general, the assessment interview has characteristics of both and might be called a semistructured instrument.

Illustration. A student requests credit for his prior experience as a radar technician in the Air Force. The assessor interviews the student and asks questions concerning his job duties and what he thinks he learned while working as a technician. Scale items are used to assess his attitudes toward his supervisors and co-workers. Open-ended questions are used to obtain information concerning his satisfaction with the job and what learning outcomes were achieved that would transfer to his present career choice or course of study. The interviewer records the student's responses, then compares this information with references, anecdotal reports, and ratings obtained from the student's supervisor in the service to arrive at an overall rating for the student's experience.

Panel Interview

This is an interviewing method whereby several interviewers jointly interview a candidate as a panel rather than through individual interviews. The panel interview is a practical technique in that it provides the opportunity for the candidate to respond to a group situation, and it provides the interviewers the opportunity to have one person interview while the others observe or write down comments, with none of the panel having to do all the observing or all the interrogating. The panel method takes less of the candidate's time since he has to answer specific questions only once rather than several times. The members of the panel should represent different ages, educational backgrounds, and interests. In addition, the interviewers should agree in advance on their roles and what points they will emphasize or questions they will ask.

Illustration. A student has contracted for a cross-cultural learning experience in which she works on a kibbutz in Israel. Her learning experience is supervised and directed by an interdisciplinary committee at the institution in which she is enrolled. Upon her return to the United States, she is interviewed by a faculty panel, the members of which represent the following subject areas: international relations, economics, Hebrew, anthropology, and social psychology. Each participant on the panel asks a sequence of open-ended questions that require the student to relate her experience to that particular discipline. After her interview, the panel evaluates her responses.

Oral Test

When assessment is desired for those work experiences that demand a high proficiency in dealing personally with subordinates, associates, or the public, an oral test should be considered. It provides the assessor with the opportunity of

seeing the candidate in the give-and-take of a personal conference, but it differs from the general concept of an interview in that it is especially planned and conducted to draw out verifiable evidence of demonstrated qualifications, such as those pertaining to personality and experience. It is especially staged to ask more searching questions than those contained in a more superficial assessment interview in which, perhaps, the candidate has merely to list and describe briefly the functions he or she has performed on the job. It is during this procedure that the assessor can judge the candidate's command of language, particularly the fluency and flavor of his or her oral vocabulary as contrasted with his or her written vocabulary. In addition, the following dimensions can be assessed with this technique: voice, appearance, alertness, ability to present ideas, poise, bearing, tact, and judgment.

One form of the oral test is the group oral performance test, in which several applicants are seen at one time and interact with each other and with the interviewer or a panel. In such a group interview, several students can respond directly to questions or problems posed by the interviewer and to each other, which makes it very similar to an academic seminar. When all the interaction is between the interviewees and when one realistic problem concerning work experience is discussed, the exercise takes on the appearance of a leaderless group discussion.

Illustration. A candidate has spent the previous seven years running a household and rearing two children. Prior to marriage, she attended a liberal arts college for two years. During her marriage, she has organized and directed children's play groups with other mothers. In addition, she has spent considerable time as a member of a baby-sitting pool and therefore has had experience in taking care of children other than her own. For assessment, she is required to participate in a group interview with three other women with similar backgrounds. In the interview the women must develop a list of skills or dimensions relevant to their home management and child-rearing experiences in order to rate themselves individually on these dimensions. The interviewer will act as a facilitator by asking appropriate questions to keep the discussion focused on the skills inventory and by probing the individuals when more information is needed.

The Oral Report or Presentation

The oral report can be best described as a combination of the essay technique and a product assessment. For a more detailed explanation, see the sections of this report concerning essay examinations and product assessment.

Oral Trade Tests

These tests contain a short series of questions about specialized trade knowledge. The tests are developed by formulating questions from information gathered through direct observation of jobs and consultations with foremen and highly skilled workers in a particular trade. The questions are revised after tryouts with skilled workers in each type of job. The questions are written so as to be fairly

easy for anyone who has actually worked in a particular job, but unfamiliar to persons who have never worked in a particular trade. Such tests are often used as interview aids in personnel selection along with other questions concerning personal qualifications. The United States Employment Service (USES) has many series of oral trade questions that cover more than 250 occupations, including those of carpenter, paper hanger, plasterer, sheet metal worker, and bricklayer.

Illustration. An engineering student wishes to receive credit for his experience as a self-employed electrical contractor. He is interviewed by an engineering professor. The interview consists of a set of 15 questions taken from the USES lists. In addition to the oral trade test, the assessor asks questions that require the candidate to relate his work experience to his course in electrical engineering.

References

Bingham, W., and Moore, B., *How to Interview*. New York: Harper, 1959.

A simply written guide to constructing and performing the interview, this book contains an entire chapter on oral examinations.

Carter, H.D., "How reliable are good oral examinations?" *California Journal of Education*, 1972, 13, 147-153.

To ascertain the reliability of the oral exam, the investigator analyzed data from a sample of 250 candidates examined in a field of medical specialization. The results indicate that the oral exam is highly reliable and that there is a low relationship between the oral exam and a written exam covering the same subject area.

Mandell, M.M., *The Selection Process: Choosing the Right Man for the Job*. New York: The American Management Association, Inc., 1964.

Although this book emphasizes techniques concerning the hiring of individuals in 8 different occupational groups, Chapters 12, 13, and 14 are particularly informative on the subject of interviewing, and much of the material can be interpreted in the light of assessment rather than selection. The chapters contain descriptions of the stress interview and the group oral examination.

Platt, J.E., "On maximizing the information obtained from science examinations—written and oral." *American Journal of Physics*, 1961, 29, 111-122.

The author regards the written examination as a mapping function in which the student's abilities are mapped on a numerical scale. An oral exam is described as a search path to locate the boundaries of a student's knowledge. Common pitfalls of the oral exam are described.

VIII. SELF-ASSESSMENT

Self-assessment or self-evaluation can be regarded as a set of relatively undefined techniques by which the student judges his own level of accomplishment. Since the student is likely to be knowledgeable about his past accomplishments, it is reasonable to assume that with appropriate guidance he can provide accurate indication of his present competences. Self-assessment can cover broad areas such as attitudes or values or it can focus on specific knowledge or skills.

At a very informal level, self-assessment consists of discussions between the student and faculty advisor to determine whether predefined learning goals are being accomplished. A somewhat more formal approach might be the preparation of a diary or report by the student on off-campus learning experiences and an evaluation of these experiences, in relation to goals, by faculty advisors. On the most formal level, self-assessment is conducted through the completion of structured checklists, occupational history information blanks, and self-assessment tests.

A *job inventory checklist* is a comprehensive listing of typical or critical tasks conducted by those engaged in an occupation. The student is asked to indicate whether or not he can perform a given task and, in some instances, how well he can do it. For example, in response to an item such as "maintenance and repair of high voltage switchgear," the student with experience in electrical technology might give any one of the following responses:

1) I know little or nothing about this.
2) I have had study or training in this.
3) I have done this under close supervision.
4) I have done this on my own under normal supervision.
5) I am consulted by other workers who have difficulty with this, or I am called to do unusually difficult jobs.

Various ways of scoring the checklist are possible. Some of them account for the differential difficulty and complexity of the listed tasks.

The *occupational history information blank* resembles a job application blank in that it seeks information about relevant previous training, employment, duties and responsibilities, community activities, and areas of special competency. It is a useful supplement to the job checklist because it can be used to verify claims made on it. Although the blank is not as highly structured as the checklist, it provides a useful format for recording relevant information in a systematic way.

Self-assessment tests are generally objective written examinations in multiple-choice format. They differ from other subject matter achievement tests only in the way they are used. Several professions, mainly in the health field, make use of such tests. The practitioner can assess his knowledge of a field by receiving feedback on his test performance and can compare his knowledge to that of other practitioners without the threat of being evaluated. He can plan his continuing education in light of the strengths and weaknesses indicated by his test scores.

Illustration. A person who repairs cars as a hobby wants to enroll in his local junior college in a program leading to an associate degree in automotive mechanics. He completes a job inventory checklist that describes in detail the activities involved in the occupation, then reviews his responses with his advisor and compares them to the responses of master mechanics on which the checklist was normed. This procedure forms a basis for a decision as to which technical courses the student can bypass and which he must take in order to gain full technical competence.

Development. If self-assessment is to be successful as an evaluation technique, the student must be given guidance in how to do it. Since he is not likely to be an objective evaluator of his own accomplishments, he must be provided with specific standards and criteria. In college-sponsored programs, such evaluations can be made in relation to a predetermined set of goals.

The development of a job inventory checklist must start with an identification of the critical tasks of the job. The checklist items should describe a particular task or skill in fairly concrete terms so that the student will have a benchmark for his performance. The development of the occupational history information blank involves the establishment of a standardized format for documenting the nature and extent of previous work experiences and their relation to academic courses.

Technical considerations. Job inventory checklists and occupational history blanks should be normed and validated prior to their use for administrative purposes. The instruments might be administered to a group of persons who are known to be competent in a particular subject matter and their responses used as a yardstick for evaluating student responses. The problem of falsification can be minimized by using a combination of both instruments and by checking the claims of students with their previous job supervisors.

Evaluation. Although the informal methods of self-assessment (for example, periodic discussions between a student and his mentor) have the appeal of flexibility and individuality, their subjective nature and lack of uniformity make them difficult to evaluate from a psychometric viewpoint. In addition to its assessment function, realistic self-appraisal is useful both for the identification and improvement of weak skills and for motivating learning. To a large degree the accuracy and the utility of the self-assessment process will depend on the effort that student and mentor expend on it and the amount of guidance or training that the student receives in the process.

Both the job inventory checklist and the occupational history information blank show promise in the assessment of different kinds of experiential learning. Since these instruments have been employed primarily in industrial settings, additional research is needed to determine whether they are applicable to higher education.

References

Primoff, E.S., *Research on Efficient Methods in Job Element Examining, Report No. 1—Research of the Additive Checklist*. Washington, DC: U.S. Civil Service Commission, 1970. (Unpublished study.)

This study describes how to develop a job inventory checklist and reports the results of studies using the checklist for the selection of workers in various trades.

Job Activity Self-Report Questionnaire (JASRQ), Lafayette, IN: Purdue Research Foundation, 1969.

This questionnaire includes a listing of various job activities and conditions. For each one the candidate is asked to indicate how much or what type of experience he has had, his present level of ability, and the level that he might be able to achieve with additional experience.

IX. RATINGS

Ratings are procedures for quantifying the opinions of one person about the behavior of another. They are judgments of the knowledges, skills, abilities, traits, or behaviors of an individual. Although they are frequently quantitative statements, they are subject to all the characteristic errors of human judgment and therefore cannot be regarded as objective descriptions of fact. Ratings are generally employed in situations where it is difficult to obtain more objective measures of an individual's competence.

Numerous rating methods exist but most can be classified into ranking, rating scale, and checklist procedures. In the ranking procedure, the rater ranks or orders the individuals in a given group from best to poorest on a specific dimension of behavior. The rating-scale methods provide a scale for estimating levels of performance, either in discrete categories or in a graphic or continuous form. The rater checks the category that he feels best describes the person being rated. The checklist methods provide the rater with a list of prescaled descriptions of behavior and the rater checks the items that best describe the person being rated.

Illustration. A student works in a college-sponsored program as an aide in a community day care center under the direction of professional day care personnel. At the conclusion of the semester, the student's field supervisor and on-site supervisor complete a rating form to evaluate the student's skills on a number of different job dimensions such as working relationships with children, planning and organizing the work, and written and oral expression.

Development. The factors or dimensions used in ratings should be relevant to the job or the subject matter being evaluated. Too many dimensions will confuse the rater, but too few are not likely to provide an adequate description of performance. The factors used should be defined as precisely as possible and should be included in the rating only if the rater can observe or has knowledge of the trait or behavior described by the factor.

The effort required to develop ratings varies with the procedure. The ranking methods are the simplest to develop and use, and checklist scales the most cumbersome because a large number of statements representing various levels of competency must be assigned scale values through fairly elaborate scaling techniques. A rating form having between 5 and 15 scales (each assessing a different dimension) is probably suitable for many purposes. Each scale point should have a specific behavioral description attached to it. A 5- or 7-point scale is likely to give an adequate discrimination of performance in most cases. Generally, raters do not have the capacity to discriminate among too many levels (for example, 20). In summing the scores on the individual scales to obtain a total score, the scales measuring the important factors should be given the most weight.

Technical considerations. The reliability and validity of ratings is a function of the rater, the rating method, and the behavior being rated. Since ratings are subjective estimates of behavior, complete accuracy is seldom possible. Following are some of the major errors or biases that tend to reduce the accuracy of ratings:

1) the tendency to rate too leniently or too severely,
2) the tendency to avoid the extremes of the scale and to rate at the average,
3) allowing an outstanding or inferior trait or aspect of performance to influence the rating of other factors ("halo effect"),
4) judging the ratee according to a personal stereotype,
5) the tendency to prejudge the ratee by an initial impression rather than on the basis of observed performance.

Accurate ratings are likely to result if the rater is trained in the rating process, is familiar with the behavior and characteristics of the ratee, and takes his time in making the ratings. Some of the rating techniques are more likely to control certain rater errors than others. For example, rankings and forced-choice checklists are generally the most successful in minimizing the leniency error prevalent in many work-appraisal situations. Ratings of simple and tangible behavior are more likely to be accurate than ratings of complex behavior. It is much easier to judge an individual's proficiency in a foreign language than to judge his appreciation and understanding of a foreign culture. The more persons who rate an individual, the more reliable the pooled ratings are likely to be. Thus, it is useful to involve the ratee's peers and clients, if any, in addition to his supervisor in the rating process.

Evaluation. Despite the many human errors and biases associated with ratings, they are likely to be indispensable in assessing those kinds of learning and accomplishment that do not lend themselves to more objective measurement. Since each of the rating techniques has certain advantages and disadvantages, a choice of which to use will depend on the situation. The ranking procedure is a simple and natural technique, but it is difficult to use with large groups and it provides only normative comparisions between individuals. The rating-scale methods vary greatly in the quality and are quite susceptible to many of the rating errors. The checklist methods are relatively resistant to bias but are difficult and cumbersome to develop.

References

Ghiselli, E.E., and Brown, C.W., *Personnel and Industrial Psychology*. New York: McGraw-Hill, 1955, 87-125.

This book provides a taxonomy of rating methods with illustrations and describes the factors that affect reliability and validity of ratings.

Mandell, M.M., *The Selection Process: Choosing the Right Man for the Job*. New York: The American Management Association, Inc., 1964, 255-284.

The author describes the rating process in the context of reference checking when selecting applicants for jobs. He identifies the problems that may arise with different types of raters and describes some of the practical aspects of evaluating a person's background through rating scales.

X. PRODUCT ASSESSMENT

Product assessment, as used in the context of experiential learning, is generally taken to mean the evaluation of a variety of student products such as reports, diaries, poems, musical compositions, drawings, paintings, and photographs. Since each of these products is assumed to represent the skill of the student in a particular activity, product assessment is an indirect procedure for evaluation of the relevant skills that students may learn in off-campus settings.

The need to evaluate products stems from the evaluator's inability to observe the performance of the student while he is creating it. Performances resulting in some types of products, such as those in creative writing, are inherently unobservable, while other kinds of performances, such as those that occur in off-campus locations or prior to the student's enrolling in the college, are not readily accessible to evaluation because of their remoteness.

The actual evaluation of a product is generally conducted by one or more experts in a particular subject matter. Since many of the products in a student's portfolio are primarily aesthetic in nature (for instance, a painting or a musical composition), subjective judgment must be made on their quality. With works such as diaries and reports, the value of the experiences described must be judged rather than the actual product.

Illustration. A student spends a year in a foreign country as part of a cultural-exchange program. He studies the history, art, and culture of the host country by visiting museums, art galleries, and architectural landmarks. After his return to the sponsoring college, he submits a diary of his experiences for evaluation to a committee of three faculty members. The committee asks him to write essays on several specific topics that will demonstrate his understanding of the foreign culture. The diary and the essays are rated individually by the committee members on a quantitative rating scale. The three sets of ratings are pooled and the summary ratings provide the basis for judging the student's appreciation of the foreign culture.

Development. A decision must be made on whether any product that the student submits is to be considered or whether the evaluation will be limited to certain kinds of products. When evaluating an art portfolio, for example, three or four kinds of media, such as drawings, paintings, and so on, can be requested of the student. If product evaluation is to be conducted in a systematic and objective manner, standards or criteria should be established prior to the evaluation process. The assessment can be facilitated and made more objective by the use of rating scales (see the section on ratings, beginning on page 47).

Technical considerations. One advantage of structuring a portfolio is that the assessment procedure can be made uniform for all students. This advantage, however, is trivial when the number of students being assessed in a particular field is very small. A disadvantage of specifying or limiting the products to be evaluated is the loss of flexibility that results in not considering the works of students who do not "fit the mold."

A problem of evaluating diaries, reports, and other descriptions of experiences is that students are likely to place themselves in a favorable light and exaggerate their accomplishments. It is possible, of course, to verify the accuracy of reports by requesting the student to provide supporting evidence, such as official documents, publications, and professional certification. Evaluations and verifications should also be sought from the student's previous supervisors and colleagues who may have participated in his learning experiences.

Evaluation. Product assessment is likely to be appropriate when

1) the product is more important than the performance leading to the product,
2) the performance leading to the product cannot be observed,
3) the product can be measured accurately and objectively.

Product assessment is especially appropriate to experiential learning because such learning necessarily takes place outside the college classroom and can seldom be observed. Although it is not as costly as performance testing and some of the other assessment techniques, if done properly it can be time-consuming. Furthermore, in most cases the judgment of several faculty members or other experts is required to make the process a success.

Reference

Fitzpatrick, R., and Morrison, E.J., "Performance and Product Evaluation," in Thorndike, R.L. ed., *Educational Measurement* (2nd Ed.). Washington, DC: American Council on Education, 1971, 261-262.

These pages discuss the issues that need to be considered when deciding whether to evaluate performance or product.